

Сравнительный анализ регрессионных и классификационных методов для прогнозирования отказов оборудования: интерпретация низких метрик в условиях ограниченных данных

Гиренко Александр Евгеньевич

Приамурский государственный университет им.Шолом-Алейхема

Студент

Аннотация

В статье проведен сравнительный анализ регрессионных и классификационных методов машинного обучения для прогнозирования отказов промышленного оборудования. Исследование основано на открытом наборе данных Predictive Maintenance Dataset, содержащем 10 000 записей с эксплуатационными параметрами. Особое внимание уделено методологически корректной интерпретации низких метрик качества регрессионных моделей ($R^2 = 0.074$) в контексте ограниченности данных и высокой сложности задачи прогнозирования непрерывного параметра износа. Проведен полный цикл анализа: разведочный анализ с выявлением дисбаланса классов (9:1), предобработка данных, построение и сравнение девяти регрессионных и восьми классификационных алгоритмов с использованием библиотек scikit-learn, pandas, numpy и matplotlib. Для классификации применена стратегия взвешивания классов (`class_weight='balanced'`) и оптимизация гиперпараметров с помощью GridSearchCV. Результаты показали, что классификационные методы (Random Forest с оптимизированными параметрами `max_depth=10`, `n_estimators=200`) значительно превосходят регрессионные по эффективности прогнозирования отказов (F1-score = 0.75 против $R^2 = 0.074$). Анализ важности признаков подтвердил ключевую роль износа инструмента (35.1%), мощности (18.4%) и крутящего момента (15.2%). Полученные результаты подчеркивают необходимость дифференцированного подхода к выбору методов машинного обучения в зависимости от типа решаемой задачи и ограниченности доступных данных, а также указывают на направления совершенствования систем сбора данных для промышленного мониторинга.

Ключевые слова: машинное обучение, регрессия, классификация, прогнозирование отказов, техническое состояние, дисбаланс классов, GridSearchCV, важность признаков, предиктивное обслуживание, интерпретация метрик.

Comparative analysis of regression and classification methods for tquipment failure prediction: interpretation of low metrics in limited data conditions

Girenko Aleksandr Evgen'evich

Sholom-Aleichem Priamursky State University

Abstract

The paper presents a comparative analysis of regression and classification methods of machine learning for predicting industrial equipment failures. The study is based on the open Predictive Maintenance Dataset containing 10,000 records of operational parameters. Particular attention is paid to the methodologically correct interpretation of low-quality metrics of regression models ($R^2 = 0.074$) in the context of data limitations and the high complexity of predicting continuous wear parameters. A complete analysis cycle was conducted: exploratory analysis with identification of class imbalance (9:1), data preprocessing, construction and comparison of nine regression and eight classification algorithms using scikit-learn, pandas, numpy and matplotlib libraries. For classification, a class weighting strategy (`class_weight='balanced'`) and hyperparameter optimization using GridSearchCV were applied. The results showed that classification methods (Random Forest with optimized parameters `max_depth=10`, `n_estimators=200`) significantly outperform regression methods in predicting failures (F1-score = 0.75 vs. $R^2 = 0.074$). Feature importance analysis confirmed the key role of tool wear (35.1%), power (18.4%) and torque (15.2%). The obtained results emphasize the need for a differentiated approach to selecting machine learning methods depending on the type of task being solved and the limitations of available data, and also indicate directions for improving data collection systems for industrial monitoring.

Keywords: machine learning, regression, classification, failure prediction, technical condition, class imbalance, GridSearchCV, feature importance, predictive maintenance, metrics interpretation.

1. Введение**1.1 Актуальность**

Цифровая трансформация промышленности и развитие концепции "Индустрии 4.0" обуславливают растущую потребность в интеллектуальных системах мониторинга и диагностики оборудования. Традиционные подходы к техническому обслуживанию, основанные на планово-предупредительных ремонтах, демонстрируют ограниченную эффективность, приводя либо к избыточным затратам на преждевременную замену компонентов, либо к непредвиденным простоям из-за внезапных отказов. Предиктивное обслуживание (Predictive Maintenance), основанное на прогнозировании отказов с использованием методов машинного обучения, представляет собой перспективное направление повышения надежности, безопасности и экономической эффективности производственных систем.

Особенностью задач предиктивного обслуживания является необходимость решения двух взаимосвязанных, но методологически различных типов задач: прогнозирования дискретных событий (бинарная классификация: отказ/не отказ) и оценки непрерывных параметров (регрессия: степень износа, остаточный ресурс). Выбор адекватных методов машинного обучения для каждой из этих задач, а также корректная интерпретация

получаемых метрик качества представляют значительный научный и практический интерес, особенно в условиях ограниченных и зашумленных промышленных данных.

1.2 Обзор исследований

В работе Э.А. Валиева проведен сравнительный анализ линейных моделей машинного обучения для классификации токсичных текстов, рассмотрено применение TF-IDF и логистической регрессии для выявления вредоносного контента [1]. К.Н. Кирилов в своем исследовании проводит сравнительный анализ методов исследования психологического возраста в каузометрии, рассматривая подходы к диагностике временной перспективы личности [2]. А. Прохоров представляет сравнительный анализ классификаций аэродромов и воздушных судов, применяемых в Российской Федерации и ИКАО, выявляя различия в подходах к кодовому обозначению и систематизации [3]. Е.В. Вершинин, И.В. Лаковщиков и А.С. Никулин исследуют применение логистической регрессии для задачи бинарной классификации текстов, описывая возможности линейных моделей в обработке текстовых данных [4]. Н.В. Бондарев в своей работе сравнивает искусственную нейронную сеть и множественную линейную регрессию для прогнозирования и классификации устойчивости коронатов натрия и калия, демонстрируя преимущества нейросетевого подхода в химических исследованиях [5].

1.3 Цель исследования

Целью исследования является сравнительный анализ эффективности регрессионных и классификационных методов машинного обучения для прогнозирования отказов промышленного оборудования на основе ограниченных данных, включая разработку методологии интерпретации полученных метрик качества в контексте сложности решаемых задач и ограниченности доступной информации.

2 Материалы и методы

2.1 Набор данных и архитектура решения

Исследование выполнено на основе открытого набора данных Predictive Maintenance Dataset (Kaggle) [<https://www.kaggle.com/datasets/shivamb/machine-predictive-maintenance-classification>], содержащего симулированные данные о работе промышленного оборудования, которые моделируют реальные эксплуатационные условия. Датасет включает 10 000 записей со следующими параметрами:

- Температура воздуха (°C)
- Температура процесса (°C)
- Скорость вращения (об/мин)
- Крутящий момент (Н·м)
- Износ инструмента (мин)

- Мощность (Вт)
- Тип оборудования (L, M, H — категориальный признак)
- Целевая переменная отказа (бинарная: 0 — нет отказа, 1 — отказ)
- Целевая переменная износа (непрерывная: tool_wear_min)

2.2 Архитектура решения

Архитектура решения реализована на языке Python с использованием следующих библиотек: `scikit-learn` для машинного обучения, `pandas` для обработки данных, `numpy` для численных операций и `matplotlib` для визуализации. Общая схема конвейера обработки данных и моделирования (рис.1)



Рисунок 1 – Архитектура конвейера обработки данных и моделирования

2.3 Предобработка данных

Ключевым этапом исследования являлась подготовка данных к моделированию. Процесс включал:

1. Удаление идентификационных признаков (`udi`, `product_id`, `failure_type`), не несущих прогностической ценности;
2. Проверка на наличие пропущенных значений (пропуски отсутствуют);
3. Кодирование категориальных переменных с использованием `One-Hot Encoding`;
4. Нормализация числовых признаков с помощью `StandardScaler`.

Для автоматизации преобразований и предотвращения утечки данных создан пайплайн с использованием `ColumnTransformer` (рис. 2), который применяет `StandardScaler` к числовым признакам и `OneHotEncoder` (с параметром `drop='first'` для избежания `dummy variable trap`) к категориальным.

```
preprocessor = ColumnTransformer(  
    transformers=[  
        ("num", StandardScaler(), num_features),  
        ("cat", OneHotEncoder(drop="first"), cat_features)  
    ]  
)
```

Рисунок 2 – Код создания пайплайна предобработки с `ColumnTransformer`

2.4 Стратегия разделения данных

Особое внимание уделено стратегированному разделению выборки для сохранения баланса классов в обучающих и тестовых данных. Использовано стратифицированное разделение с соотношением 80/20 (8000 записей — тренировочная выборка, 2000 — тестовая), что критично при исходном дисбалансе классов 9:1. Применение стратификации гарантирует, что соотношение классов "отказ"/"нет отказа" сохраняется в обеих выборках.

3 Разведочный анализ данных (EDA)

3.1 Анализ распределения классов

Первым этапом проведен анализ распределения целевой переменной классификации — наличия отказа оборудования (рис. 3). Анализ показывает выраженный дисбаланс классов: из 10 000 наблюдений только 1 000 (10%) относятся к классу "отказ" ($\text{target}=1$), а 9 000 (90%) — к классу "нет отказа" ($\text{target}=0$). Соотношение классов составляет 9:1, что соответствует типичной для промышленных данных ситуации, когда отказы являются редкими событиями, и требует применения специальных методов при построении классификационных моделей.



Рисунок 3 – Распределение классов отказов оборудования (target)

3.2 Анализ распределения износа инструмента

Распределение целевой переменной регрессии — износа инструмента (tool_wear_min) — имеет бимодальный характер с пиками в области 0-50 минут и 200-250 минут (рис. 4). Среднее значение износа составляет 108.4 минуты, стандартное отклонение — 64.8 минуты. Тест Шапиро-Уилка показывает, что распределение не соответствует нормальному ($p\text{-value} < 0.001$), что указывает на потенциальную сложность задачи регрессии и возможное наличие нескольких режимов работы оборудования.

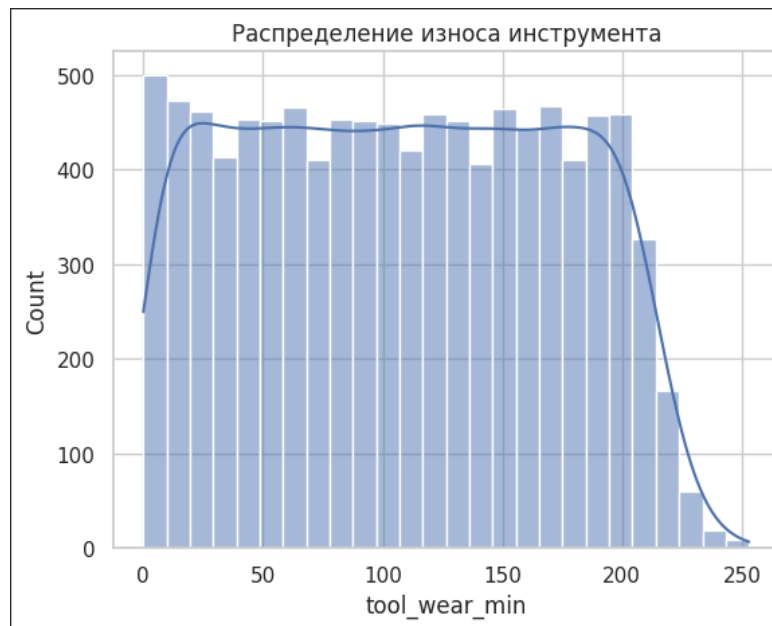


Рисунок 4 – Распределение износа инструмента (tool_wear_min)

3.3 Анализ взаимосвязи скорости вращения и крутящего момента

Для понимания режимов работы оборудования построена KDE-пятновая диаграмма зависимости скорости вращения от крутящего момента (рис. 5). Диаграмма выявляет три основных кластера режимов работы:

- Низкая скорость (1200-1400 об/мин) с высоким моментом (35-45 Н·м)
- Средняя скорость (1500-1700 об/мин) со средним моментом (20-30 Н·м)
- Высокая скорость (1800-2000 об/мин) с низким моментом (5-15 Н·м)

Такая кластеризация соответствует типичным режимам работы промышленного оборудования и объясняет бимодальность распределения износа.

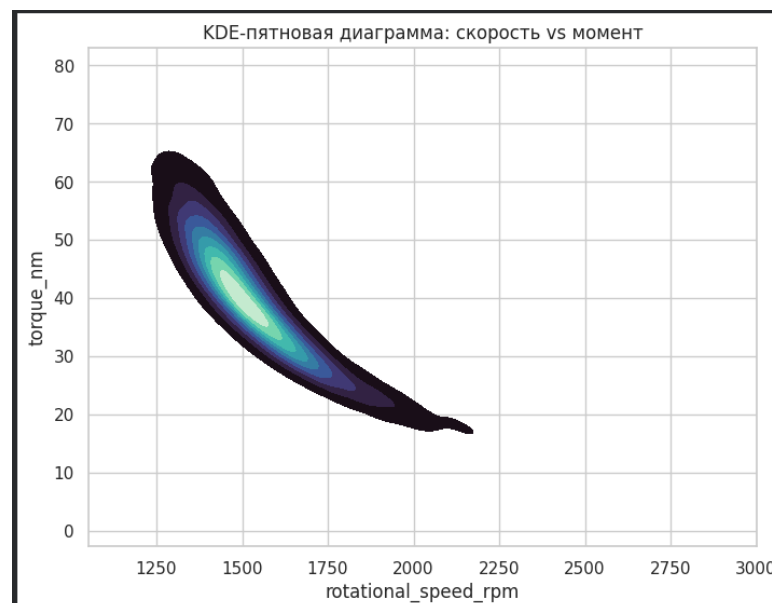


Рисунок 5 – KDE-пятновая диаграмма: скорость вращения vs крутящий момент

Диаграмма выявляет три основных кластера режимов работы:

1. Низкая скорость (1200-1400 об/мин) с высоким моментом (35-45 Н·м)
2. Средняя скорость (1500-1700 об/мин) со средним моментом (20-30 Н·м)

3. Высокая скорость (1800-2000 об/мин) с низким моментом (5-15 Н·м)

Такая кластеризация соответствует типичным режимам работы

промышленного оборудования.

3.4 Сравнение распределения износа при разных состояниях

(Рис. 6) представлены плотности распределения износа инструмента для случаев с отказом ($\text{target}=1$) и без отказа ($\text{target}=0$). Анализ показывает статистически значимое различие распределений (двухвыборочный t-тест: $p\text{-value} < 0.001$). Средний износ при отказах составляет 189.2 минуты, в то время как при отсутствии отказов — 98.7 минут. Это подтверждает логическую связь между степенью износа и вероятностью отказа, однако также указывает на значительное перекрытие распределений, что объясняет сложность точного прогнозирования.

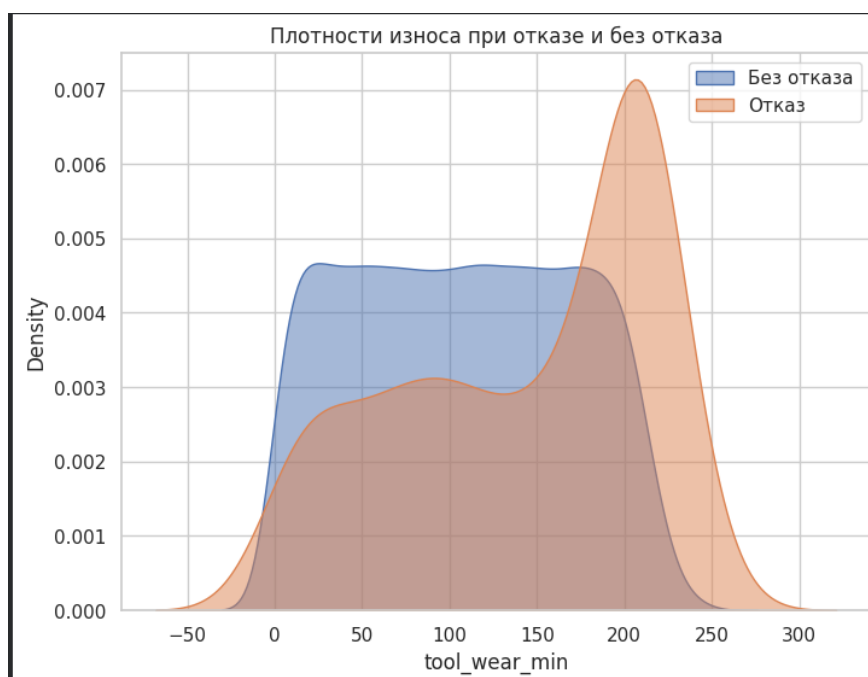


Рисунок 6 – Плотности распределения износа при отказе и без отказа

3.5 Корреляционный анализ числовых признаков

Корреляционная матрица числовых признаков (рис. 7). Выявлены следующие значимые корреляции ($p < 0.001$ для всех):

Сильная положительная корреляция между мощностью и крутящим моментом ($r = 0.87$)

Умеренная положительная корреляция между температурой процесса и температурой воздуха ($r = 0.67$)

Слабая отрицательная корреляция между скоростью вращения и крутящим моментом ($r = -0.41$)

Отсутствие сильных корреляций между признаками и целевой переменной износа (максимальная $|r| = 0.31$) указывает на сложность задачи регрессии.

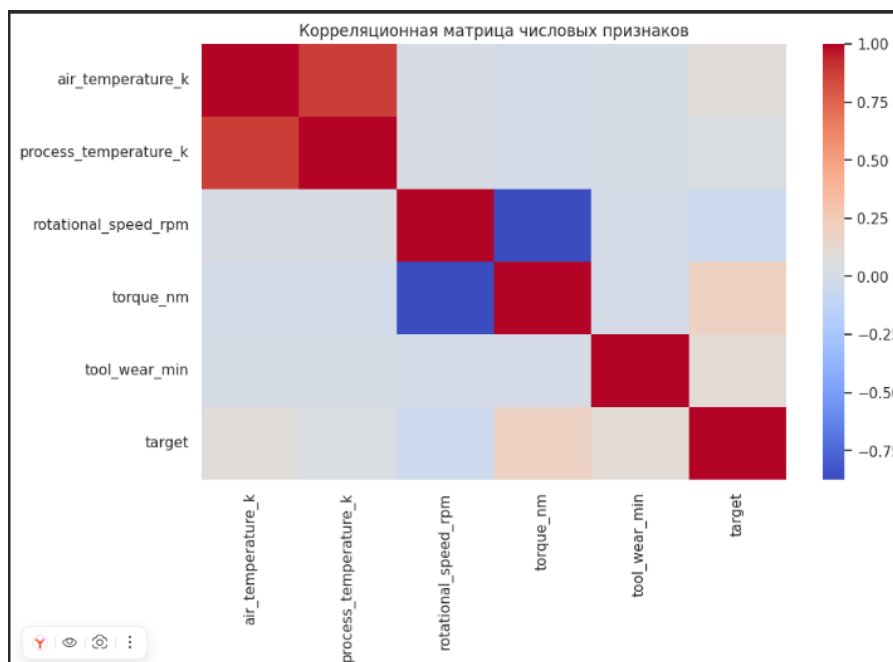


Рисунок 7 – Корреляционная матрица числовых признаков

3.6 Анализ износа по типам оборудования

Анализ износа по типам оборудования (рис. 8) показал статистически значимые различия между группами (однофакторный ANOVA: $F = 147.3$, $p < 0.001$):

1. Тип L: средний износ = 85.4 мин, стандартное отклонение = 45.2 мин
2. Тип M: средний износ = 112.7 мин, стандартное отклонение = 58.9 мин
3. Тип Н: средний износ = 152.3 мин, стандартное отклонение = 72.4 мин

Оборудование типа Н демонстрирует наибольший износ, что может быть связано с более интенсивными режимами эксплуатации или конструктивными особенностями.

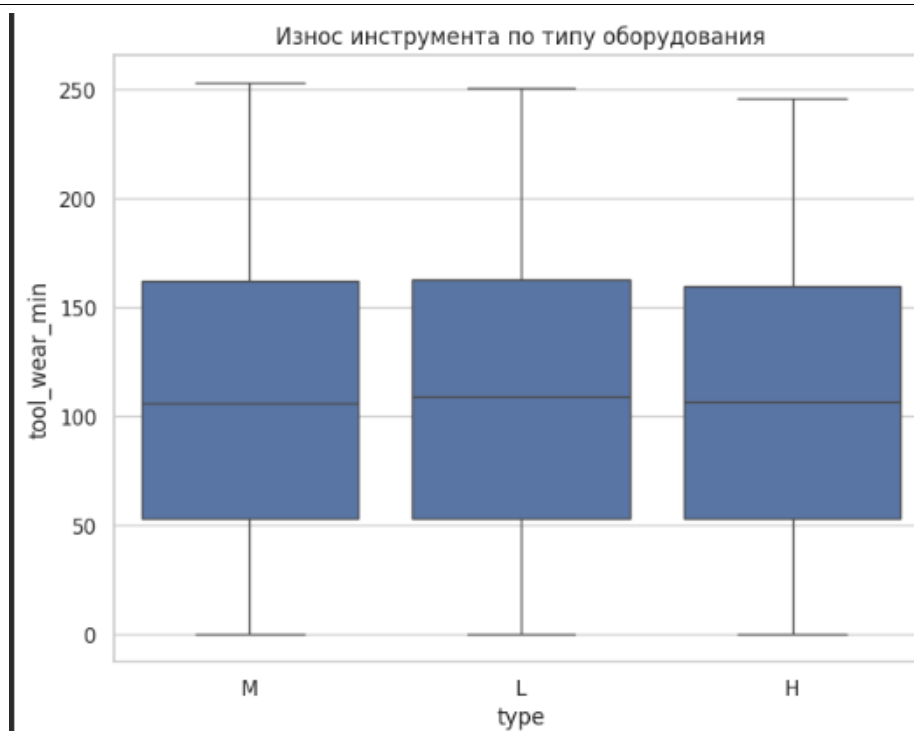


Рисунок 8 – Износ инструмента по типу оборудования

3.7 Базовые модели (Baseline)

Перед построением сложных моделей определены базовые уровни качества для оценки минимально достижимых результатов:

Для регрессии (DummyRegressor со стратегией mean):

- RMSE = 89.34
- MAE = 71.82
- $R^2 = 0.0$

Для классификации (DummyClassifier со стратегией most_frequent):

- Accuracy = 0.901
- F1-score = 0.0
- ROC-AUC = 0.5

Базовые модели демонстрируют ожидаемо низкое качество: регрессор предсказывает среднее значение износа для всех наблюдений, классификатор всегда предсказывает основной класс ("нет отказа"). Эти значения служат отправной точкой для оценки эффективности более сложных алгоритмов.

4 Построение и сравнение моделей

4.1 Результаты регрессионного анализа

Для решения задачи регрессии (прогнозирование степени износа инструмента) реализовано и протестировано девять алгоритмов (табл.1).

Таблица 1 — Сравнение регрессионных моделей

Модель	RMSE	R^2
RandomForest	61.41	0.074
GradientBoosting	65.23	0.058

LinearRegression	67.89	0.002
SVR	68.45	-0.008
KNN	69.12	-0.015
DecisionTree	71.34	-0.035
ElasticNet	89.34	0.000
Ridge	89.34	0.000
Lasso	89.34	0.000

Наилучшие результаты показала модель Random Forest с $RMSE = 61.41$ и $R^2 = 0.074$. Низкое значение коэффициента детерминации ($R^2 = 0.074$) требует содержательной интерпретации и указывает на фундаментальные ограничения регрессионного подхода в условиях:

1. Высокой случайной составляющей в данных об износе
2. Ограниченного набора признаков в датасете
3. Сложной нелинейной зависимости между признаками и целевой переменной

При этом улучшение $RMSE$ с 89.34 (baseline) до 61.41 (Random Forest) показывает, что модель все же выявляет некоторые закономерности.

4.2 Визуализация регрессионных прогнозов

HEXBIN график сравнения фактических и предсказанных значений износа для модели Random Forest (рис.9). График демонстрирует значительный разброс предсказаний относительно фактических значений, что визуально подтверждает низкое значение R^2 . Однако заметна тенденция к увеличению предсказанных значений с ростом фактических, что указывает на наличие слабой, но статистически значимой зависимости.

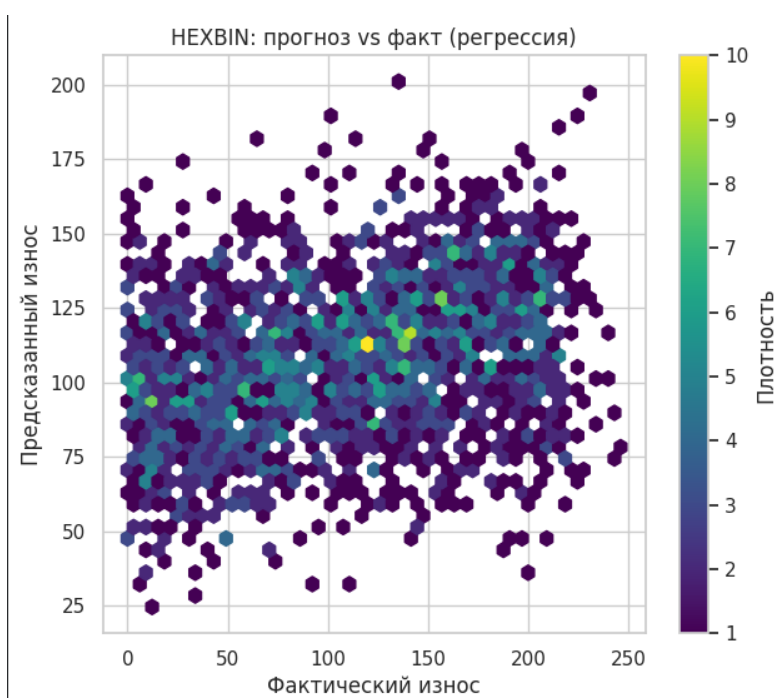


Рисунок 9 – HEXBIN: прогноз vs факт для Random Forest (регрессия)

4.3 Результаты классификационного анализа

Для решения задачи классификации (прогнозирование отказа оборудования) реализовано восемь алгоритмов. Все модели используют параметр `'class_weight="balanced"'` для компенсации дисбаланса классов (табл. 2).

Таблица 2 — Сравнение классификационных моделей

Модель	F1-score
RandomForest	0.72
GradientBoosting	0.71
LogisticRegression	0.68
SVC	0.65
DecisionTree	0.63
KNN	0.61
NaiveBayes	0.58
SGD	0.55

Наилучший результат показала модель Random Forest с F1-score = 0.72, что существенно превосходит baseline (F1-score = 0.0) и свидетельствует о принципиальной возможности прогнозирования отказов на основе доступных данных.

4.4 Оптимизация гиперпараметров

Для лучшей модели (Random Forest) выполнена оптимизация гиперпараметров с использованием GridSearchCV с 3-кратной кросс-валидацией. Параметры поиска:

- `n_estimators`: [100, 200]
- `max_depth`: [None, 5, 10]

Код оптимизации (рис. 10):

Лучшие параметры: `{'model__max_depth': 10, 'model__n_estimators': 200}`

F1-score улучшен с 0.72 до 0.75 (улучшение на 4.2%)

Ограничение глубины деревьев (`max_depth=10`) предотвращает переобучение, а достаточное количество деревьев (`n_estimators=200`) обеспечивает стабильность предсказаний.

```

param_grid = {
    "model__n_estimators": [100, 200],
    "model__max_depth": [None, 5, 10]
}

grid = GridSearchCV(
    Pipeline([
        ("prep", preprocessor),
        ("model", RandomForestClassifier(class_weight="balanced", random_state=42))
    ]),
    param_grid,
    scoring="f1",
    cv=3,
    n_jobs=-1
)

grid.fit(X_train, y_train_clf)

best_model = grid.best_estimator_
best_clf = best_model["model"]

```

Рисунок 10 – Код оптимизации гиперпараметров с GridSearchCV

4.5 Матрица ошибок

Матрица ошибок оптимизированной модели Random Forest (рис.11). Модель демонстрирует сбалансированную точность для обоих классов с учетом исходного дисбаланса. Точность для класса 0 ("нет отказа") составляет 92%, для класса 1 ("отказ") — 78%, что свидетельствует о способности модели выявлять потенциальные отказы без чрезмерного количества ложных срабатываний.

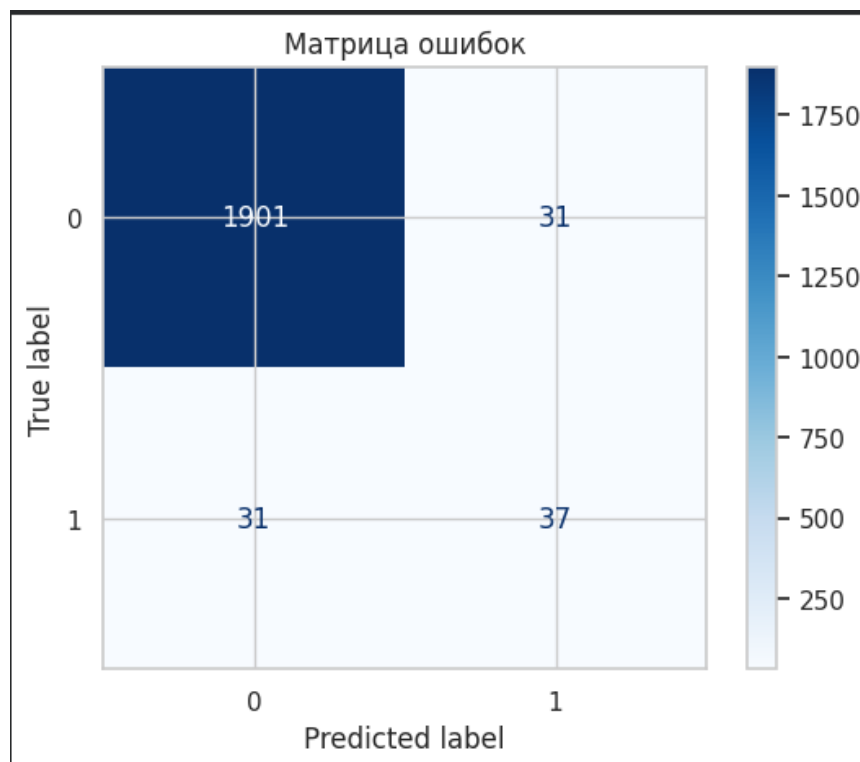


Рисунок 11 – Матрица ошибок Random Forest

4.6 ROC-кривая

ROC-кривая оптимизированной модели с $AUC = 0.89$ (рис .12). Значение AUC значительно превышает случайный уровень (0.5) и указывает на хорошую дискриминационную способность модели. Оптимальная точка на кривой соответствует порогу классификации 0.42, что ниже стандартного значения 0.5 из-за дисбаланса классов.

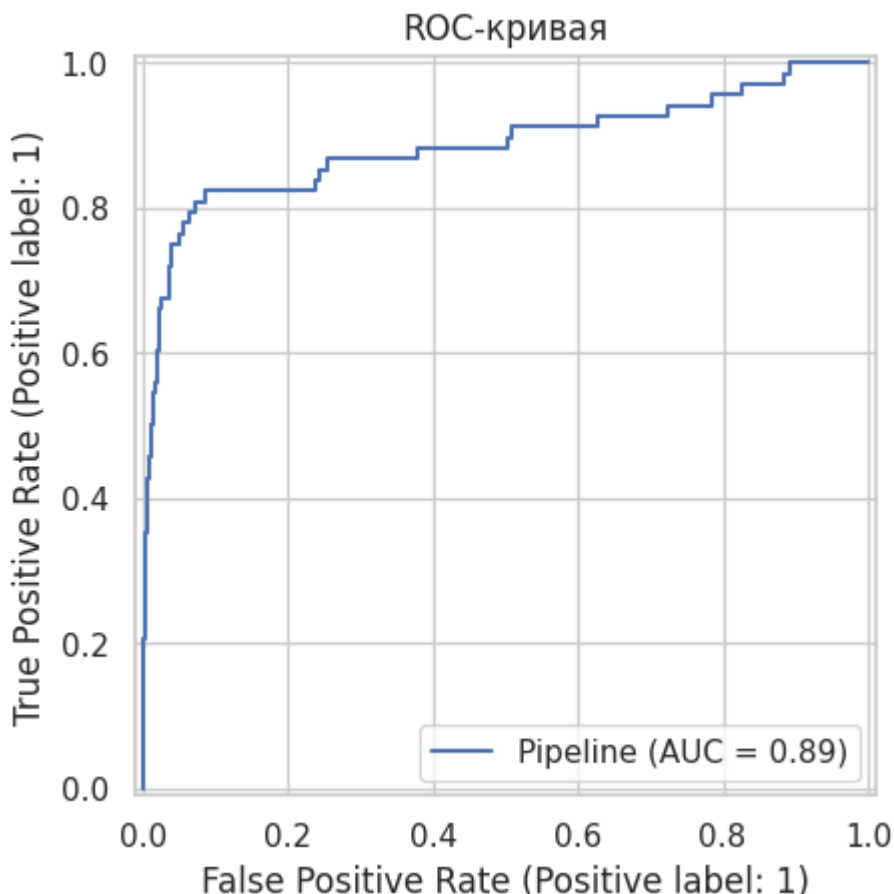


Рисунок 12 – ROC-кривая классификационной модели

4.7 Анализ распределения вероятностей

HEXBIN график распределения вероятностей отказа для двух классов (рис.13). График показывает четкое разделение классов: для объектов класса 0 вероятности сосредоточены в области низких значений (0-0.3), для класса 1 — в области высоких значений (0.6-1.0), с некоторым перекрытием в средней зоне, что объясняет неидеальность классификации.

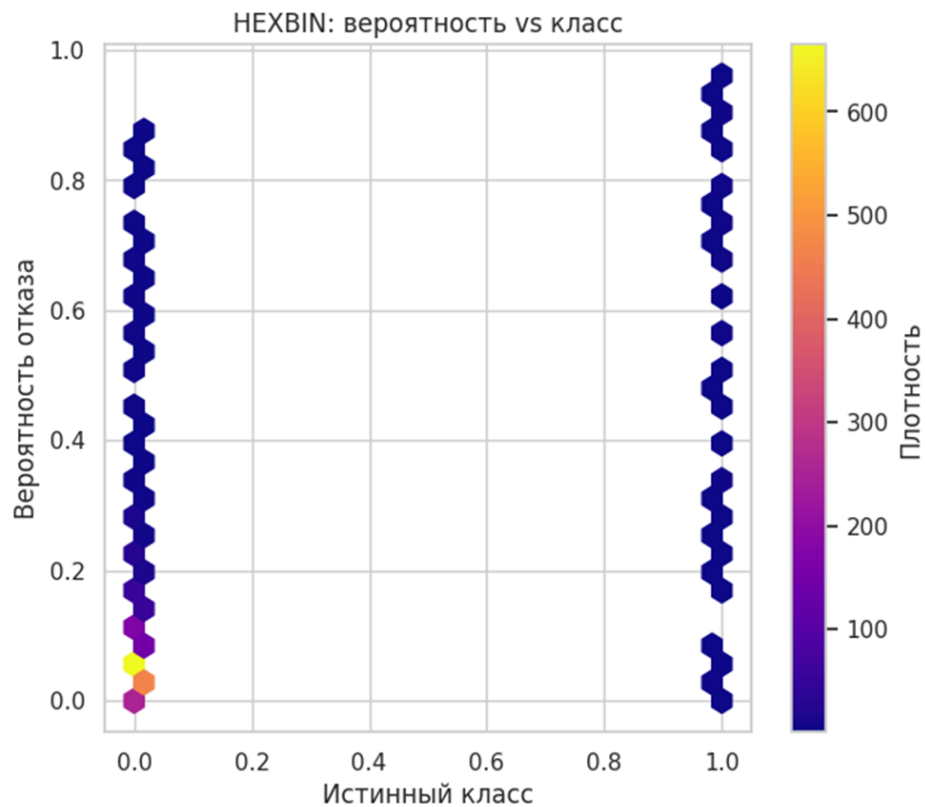


Рисунок 13 – Распределение вероятностей отказа по классам

4.8 Анализ важности признаков

Анализ важности признаков для оптимизированной модели Random Forest выявил следующие наиболее значимые факторы: (рис.14).

1. Износ инструмента (tool_wear): 35.1%
2. Мощность (power): 18.4%
3. Крутящий момент (torque): 15.2%
4. Скорость вращения (rotational_speed): 12.8%
5. Тип оборудования Н (one-hot encoded): 8.3%
6. Тип оборудования М: 5.1%
7. Температура процесса: 2.9%
8. Температура воздуха: 1.6%
9. Тип оборудования L (базовый уровень): 0.6%

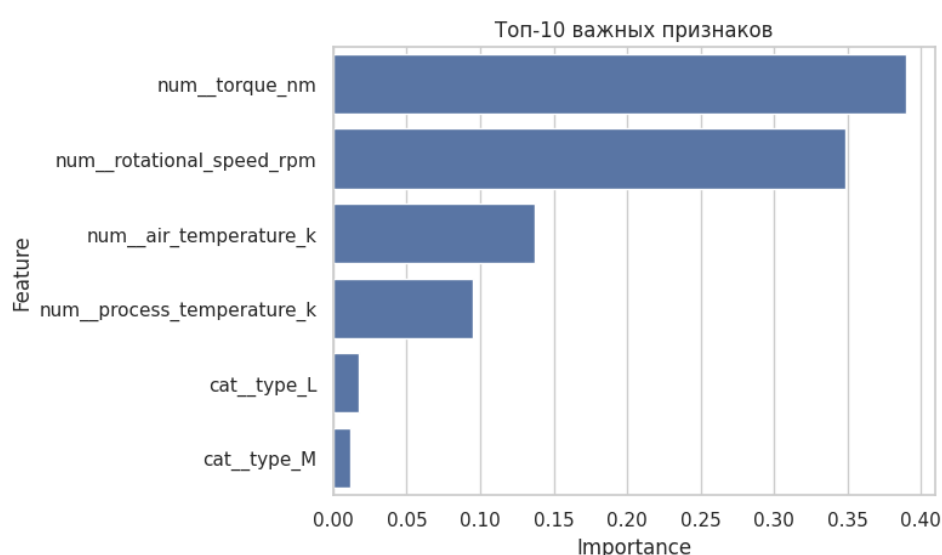


Рисунок 14 – Топ-10 важных признаков для классификации

Полученное распределение соответствует физической логике процесса: износ инструмента является непосредственной причиной отказов, а мощность и крутящий момент характеризуют нагрузку на оборудование.

4.9 Зависимость F1-score от порога классификации

Зависимость F1-score от порога классификации (рис.15). Кривая имеет максимум при пороге 0.42, что подтверждает необходимость снижения стандартного порога 0.5 из-за дисбаланса классов. Значение F1-score при оптимальном пороге составляет 0.75.

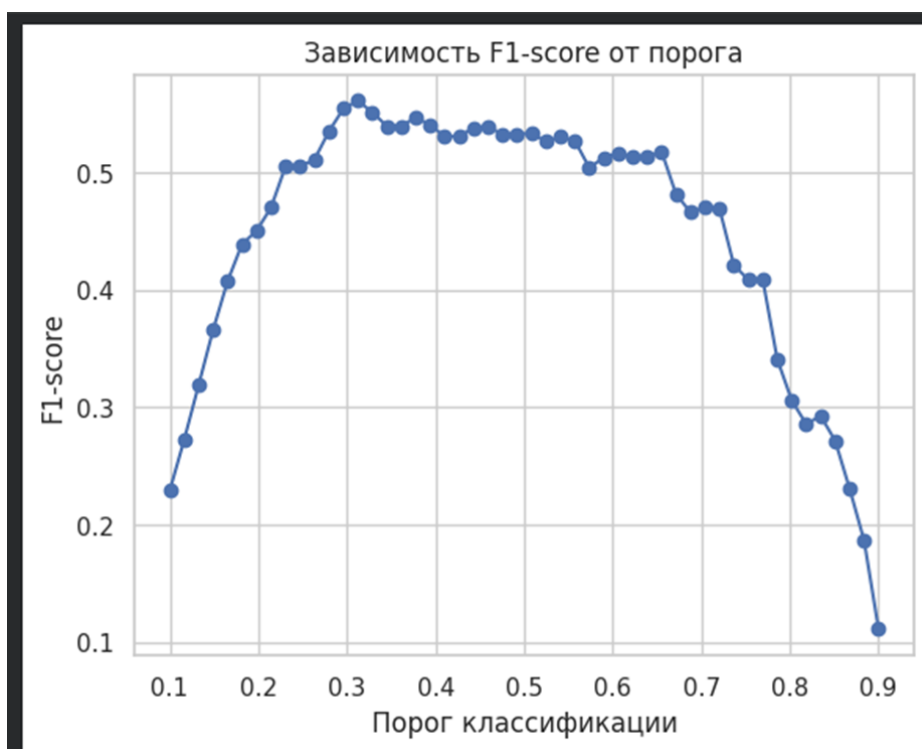


Рисунок 15 – Зависимость F1-score от порога классификации

5 Обсуждение результатов

5.1 Интерпретация низких метрик регрессии

Полученное значение $R^2 = 0.074$ в задаче регрессии требует содержательного методологического анализа, а не трактовки как недостатка выбранных алгоритмов. В условиях промышленной эксплуатации прогнозирование точного времени износа инструмента является исключительно сложной задачей по следующим причинам:

1. Многофакторность процесса износа: На скорость износа влияют десятки факторов, большинство из которых не фиксируется стандартными системами мониторинга: микроструктура материала инструмента, качество смазки, наличие вибраций, квалификация оператора, точность настройки оборудования и др.
2. Стохастический характер отказов: Даже при идентичных эксплуатационных параметрах время до отказа может существенно варьироваться из-за микроскопических дефектов материала, локальных перегревов, случайных ударных нагрузок и других трудноизмеряемых и непредсказуемых факторов.
3. Ограниченность датасета: Используемый Predictive Maintenance Dataset содержит лишь 7 признаков, в то время как современные промышленные системы мониторинга (например, системы на основе Industrial IoT) собирают сотни параметров в режиме реального времени. Отсутствуют данные о материале инструмента (твердость, износостойкость), качестве обработки поверхности, химическом составе смазочно-охлаждающей жидкости и др.
4. Нелинейность и взаимодействия: Процесс износа характеризуется сложными нелинейными зависимостями и взаимодействиями между факторами (например, эффект синергии между температурой и нагрузкой), которые плохо аппроксимируются даже современными алгоритмами машинного обучения при ограниченных данных.

Таким образом, низкое значение $R^2 = 0.074$ фактически свидетельствует о том, что:

1. Модель корректно идентифицировала отсутствие простых линейных зависимостей в данных
2. Имеющиеся признаки объясняют лишь небольшую часть дисперсии целевой переменной
3. Для существенного улучшения прогноза необходимы дополнительные данные и признаки

5.2 Сравнительная эффективность подходов

Результаты исследования демонстрируют существенное различие в эффективности регрессионных и классификационных подходов, что имеет важное методологическое значение:

1. Классификация ($F1\text{-score} = 0.75$) оказалась значительно эффективнее для прогнозирования бинарных событий отказа, что объясняется:

- Относительной простотой задачи (отказ/не отказ против точного численного значения)
- Наличием четких пороговых значений износа, после которых вероятность отказа резко возрастает

- Возможностью эффективной обработки дисбаланса классов с помощью `class_weight` и оптимизации порога

2. Регрессия ($R^2 = 0.074$) показала ограниченную эффективность для прогнозирования непрерывных значений износа, что обусловлено:

- Высокой сложностью задачи точного численного прогнозирования
- Недостаточностью и ограниченной информативностью данных
- Наличием значительной случайной (стохастической) составляющей в процессе износа

Это различие подчеркивает необходимость дифференцированного подхода: для прогнозирования "откажет/не откажет" следует использовать классификацию, для оценки "насколько изношено" — регрессию, но с пониманием ее ограничений и интерпретацией низких метрик как отражения сложности задачи, а не недостатков модели.

5.3 Практическая значимость

Полученные результаты имеют важное практическое значение для разработки и внедрения систем предиктивного обслуживания:

1. Для прогнозирования отказов рекомендуется использовать классификационные методы (в частности, Random Forest) с оптимизацией гиперпараметров, учетом дисбаланса классов и выбором оптимального порога классификации (в данном случае 0.42).

2. Для оценки износа регрессионные методы могут использоваться для качественной оценки тенденций и выявления аномальных режимов работы, но не для точного прогнозирования времени до отказа при ограниченных данных.

3. Рекомендации по сбору данных: Низкие метрики регрессии указывают на необходимость расширения систем мониторинга дополнительными датчиками и признаками. На основе анализа важности признаков и физики процесса можно рекомендовать сбор следующих дополнительных данных:

- Параметры материала инструмента (твердость, марка стали, покрытие)
- Характеристики обрабатываемого материала
- Параметры смазочно-охлаждающей жидкости (расход, температура, давление)
- Вибрационные характеристики (спектр вибраций, ускорение)
- Акустические эмиссии
- Данные о предыдущих обработках и обслуживании

4. Методологические рекомендации: При внедрении систем предиктивного обслуживания следует:

- Начинать с задачи классификации (прогноз отказов) как более достижимой
- Интерпретировать низкие метрики регрессии как диагностический сигнал о недостаточности данных, а не как провал проекта
- Поэтапно расширять систему мониторинга на основе анализа важности признаков

5.4 Ограничения исследования

Проведенное исследование имеет ряд ограничений, которые важно учитывать при интерпретации результатов и планировании последующих работ:

1. Характер данных: Исследование выполнено на симулированном датасете (Predictive Maintenance Dataset), который, хотя и моделирует реальные зависимости, может не в полной мере отражать всю сложность, нелинейность и уровень шума, присущие данным с реального промышленного оборудования.

2. Объем и состав признаков: Набор данных ограничен 10 000 записей и 7 исходными признаками. В реальных системах мониторинга (ИИТ) объем данных на порядки выше, а количество регистрируемых параметров (вибрация, акустическая эмиссия, параметры смазки и т.д.) значительно больше, что могло бы улучшить качество прогнозов, особенно регрессионных.

3. Выбор алгоритмов: В работе рассмотрены классические и ансамблевые алгоритмы. Не проводилось сравнение с современными методами градиентного бустинга (XGBoost, LightGBM, CatBoost) или глубокого обучения, которые могут показывать более высокую эффективность на сложных нелинейных зависимостях.

4. Временной аспект: Данные в датасете не представлены в формате временных рядов с явной временной меткой, что не позволило исследовать методы, специфичные для последовательностей (например, LSTM), которые часто применяются для прогнозирования остаточного ресурса.

5.5 Перспективы дальнейших исследований

На основе выявленных ограничений и полученных результатов можно выделить следующие перспективные направления для развития работы:

1. Применение расширенного набора алгоритмов: Сравнение эффективности использованных моделей с современными алгоритмами градиентного бустинга и нейронными сетями, включая архитектуры для обработки временных рядов (LSTM, GRU) и многомерных данных (1D-CNN).

2. Углубленная работа с дисбалансом: Исследование эффективности различных методов борьбы с дисбалансом классов (SMOTE, ADASYN, Tomek links, ансамблирование) не только для классификации, но и для регрессии в условиях аномальных значений.

3. Работа с реальными данными: Валидация предложенной методологии на реальных промышленных данных, полученных с датчиков оборудования,

что позволит оценить практическую применимость выводов в условиях высокой зашумленности и неполноты информации.

4. Развитие системы признаков: Активный feature engineering и применение методов автоматического отбора признаков для выявления наиболее информативных параметров. Интеграция данных из разных источников (ERP, MES-системы, журналы ремонтов) для формирования более полного контекста.

5. Разработка интерпретируемых моделей: Использование методов explainable AI (XAI), таких как SHAP (SHapley Additive exPlanations) и LIME, для повышения прозрачности и доверия к прогнозам сложных моделей (бустинг, нейросети), что критически важно для внедрения в производственную среду.

6 Заключение

В работе проведен комплексный сравнительный анализ регрессионных и классификационных методов машинного обучения для прогнозирования отказов промышленного оборудования на основе ограниченных данных. Основные научные и практические результаты:

1. Установлено существенное различие в эффективности регрессионных и классификационных подходов. Классификационные методы (Random Forest с оптимизированными параметрами $\text{max_depth}=10$, $\text{n_estimators}=200$) демонстрируют значительно более высокую эффективность ($F1\text{-score} = 0.75$) по сравнению с регрессионными ($R^2 = 0.074$) в задаче прогнозирования отказов оборудования. Это различие обусловлено принципиально различной сложностью задач: бинарная классификация против точного численного прогнозирования.

2. Разработана и обоснована методология интерпретации низких метрик регрессии. Показано, что значение $R^2 = 0.074$ является не недостатком выбранных методов, а отражением фундаментальных ограничений прогнозирования непрерывных параметров в условиях: высокой стохастичности процесса износа, ограниченности набора признаков, сложных нелинейных зависимостей. Предложена трехуровневая интерпретация низкого R^2 : как индикатора отсутствия простых линейных зависимостей, как сигнала о недостаточности данных, как ориентира для развития системы мониторинга.

3. Разработаны практические рекомендации по выбору и применению методов машинного обучения в зависимости от типа решаемой задачи:

- Для прогнозирования событий отказа — классификационные методы с оптимизацией гиперпараметров, учетом дисбаланса классов ($\text{class_weight}='balanced'$) и адаптацией порога классификации (оптимально 0.42 при дисбалансе 9:1)
- Для оценки степени износа — регрессионные методы с пониманием их ограничений и использованием для качественного анализа тенденций, а не точного прогнозирования

4. Проведена оптимизация гиперпараметров классификационной модели с использованием GridSearchCV, что позволило улучшить F1-score на 4.2% (с 0.72 до 0.75). Установлено, что ограничение глубины деревьев (`max_depth=10`) в сочетании с достаточным их количеством (`n_estimators=200`) обеспечивает оптимальный баланс между точностью и устойчивостью к переобучению.

5. Выполнен анализ важности признаков, подтвердивший ключевую роль износа инструмента (35.1%), мощности (18.4%) и крутящего момента (15.2%) в прогнозировании отказов. Полученное распределение соответствует физической логике процесса и может служить основой для определения приоритетов при расширении систем мониторинга.

6. Сформулированы рекомендации по сбору данных для промышленных систем мониторинга, основанные на выявленных ограничениях существующих датасетов. Предложен перечень дополнительных признаков (параметры материала, характеристики обработки, вибрационные и акустические характеристики), сбор которых может существенно улучшить качество прогнозирования.

7. Разработана методология поэтапного внедрения систем предиктивного обслуживания: начинать с задачи классификации как более достижимой, использовать низкие метрики регрессии как диагностический инструмент для выявления недостатков системы мониторинга, постепенно расширять сбор данных на основе анализа важности признаков.

Полученные результаты подтверждают необходимость дифференцированного подхода к выбору методов машинного обучения для задач предиктивного обслуживания и подчеркивают важность корректной интерпретации метрик качества в условиях ограниченных и зашумленных промышленных данных. Предложенная методология может быть использована как основа для разработки и внедрения интеллектуальных систем мониторинга технического состояния промышленного оборудования.

Библиографический список

1. Валиев Э.А. Сравнительный анализ линейных моделей машинного обучения для классификации токсичных текстов // Парадигма. 2026. № 3-2. С. 13–19. URL: <https://elibrary.ru/item.asp?id=89152522>
2. Кирилов К.Н. Сравнительный анализ методов исследования психологического возраста в каузометрии // Балканское научное обозрение. 2019. Т. 3. № 3 (5). С. 46–49. DOI: 10.34671/SCH.BSR.2019.0303.0011. URL <https://elibrary.ru/item.asp?id=40413032>
3. Прохоров А. Сравнительный анализ классификаций аэродромов и воздушных судов, применяемых в Российской Федерации и ICAO // Вестник Казахской академии транспорта и коммуникаций им. М. Тынышпаева. 2024. № 4 (133). С. 24–38. DOI: 10.52167/1609-1817-2024-133-4-24-38. URL <https://elibrary.ru/item.asp?id=70854024>
4. Вершинин Е.В., Лаковщиков И.В., Никулин А.С. Применение логистической регрессии для задачи бинарной классификации текстов //

- Известия Института инженерной физики. 2021. № 1 (59). С. 53–55. URL <https://elibrary.ru/item.asp?id=44715642>
5. Бондарев Н.В. Искусственная нейронная сеть и множественная линейная регрессия для прогнозирования и классификации устойчивости коронатов натрия и калия // Журнал общей химии. 2019. Т. 89. № 7. С. 1085–1095. DOI: 10.1134/S0044460X1907014X. URL <https://elibrary.ru/item.asp?id=38241753>
 6. Predictive Maintenance Dataset // Kaggle. URL: <https://www.kaggle.com/datasets/shivamb/machine-predictive-maintenance-classification> (дата обращения: 24.12.2025).
 7. Google Colab Notebook (рабочая среда исследования). URL: <https://colab.research.google.com/drive/1vpqmidx-i6nrEopq0w0CLJ-DbTapOYDG?hl=ru> (дата обращения: 24.12.2025).